This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

Spontaneous Expression Detection from 3D Dynamic Sequences by Analyzing Trajectories on Grassmann Manifolds

Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, and Stefano Berretti

Abstract—In this paper, we propose a framework for online spontaneous emotion detection, such as happiness or physical pain, from depth videos. Our approach consists on mapping the video streams onto a Grassmann manifold (i.e., space of k-dimensional linear subspaces) to form time-parameterized trajectories. To this end, depth videos are decomposed into short-time subsequences, each approximated by a k-dimensional linear subspace, which is in turn a point on the Grassmann manifold. Then, the temporal evolution of subspaces gives rise to a precise mathematical representation of trajectories on the underlying manifold. In the final step, extracted spatio-temporal features based on computing the velocity vectors along the trajectories, termed Geometric Motion History (GMH), are fed to an early event detector based on Structured Output SVM, which enables online emotion detection from partially-observed data. Experimental results obtained on the publicly available Cam3D Kinect and BP4D-spontaneous databases validate the proposed solution. The first database has served to exemplify the proposed framework using depth sequences of the upper part of the body collected using depth-consumer cameras, while the second database allowed the application of the same framework to physical pain detection from high-resolution and long 3D-face sequences.

Index Terms—Depth sequences, linear subspaces, Grassmann manifold, spontaneous expression, pain detection.

1 Introduction

With the widespread diffusion of devices endowed with on-board cameras (e.g., hand-held devices, entertainment consoles, personal computers, surveillance and monitoring sensors), there is now an increasing interest in performing online detection of spontaneous emotions and complex mental states rather than deliberate expressions. This has potential applications in the diagnostics of pathologies, such as Alzheimer and Autism, human-computer interaction, gaming, augmented and virtual reality, drivers fatigue detection and many others.

The robustness of 3D scans against illumination and pose variations in comparison to 2D images, and the recent technological advancement in 3D sensors, motivated a shift from 2D to 3D in the face analysis domain. Also, considering the spatio-temporal information (4D) data for analyzing facial expressions showed an important success in comparison to 3D static and 2D approaches [1], [2]. This trend has been further strengthened by the introduction of inexpensive acquisition devices accessible to a large number of users, such as the Kinect-like cameras that provide fast, albeit low-resolution, streams of depth data. This opened the way to new opportunities and challenges to the automatic human affect analysis.

Automatic emotion analysis is mainly guided by the discrete categorization into six basic classes, namely, *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* as proposed by Ekman [3]. Most of the benchmark public databases, such as the Cohn-Kanade [4], Multi-Pie [5] and BU-4DFE [6] datasets follow this categorization. So, several approaches present results on these datasets [7], [8]. However, deploying methods developed on such datasets in real world applications faces serious challenges because the human affect states are more complex than this six-classes representation, and spontaneous facial expressions are different from deliberate one [9]. In particular, several common affects in our daily life communication, like *confused*, *thinking*, *sadness* and *depressed* are not covered by such categorization. To describe the wide range of spontaneous affects that people show in their face to face communication, the categorization of human emotional states needs to be done in a pragmatic and context-dependent manner [10].

To address the limitations of the categorical affect description, a continuous two dimensional *arousal-valence* space has been proposed by Russell and Mehrabian [11] and Watson et al. [12]. In this space (illustrated in Fig. 1), the *valence* (horizontal) dimension measures how a human feels, from positive to negative; The *arousal* (vertical) dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive [13]. In contrast to the categorical representation, the *arousal-valence* representation enables labeling of a wider range of emotions. The automatic emotion analysis methods based on this representation tended to use binary classification between positive and negative emotions [14] or four-class classification [15]. More details on emotion representation in continuous space are given by Gunes and Björn [16].

Following the idea of detecting spontaneous emotions, other aspects are recently emerged. First, some affects have attracted particular interest, such as for *pain*. In fact, there is an approved correlation between human pain affect and cer-

- *T. Alashkar, B. Ben Amor and M. Daoudi are with Mines-Télécom/Télécom Lille; CRIStAL (UMR CNRS 9189), France. E-mail: firstname . lastname @ telecom-lille.fr*
- *S. Berretti is with Department of Information Engineering, University of Florence, Italy. E-mail: stefano.berretti@unifi.it*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

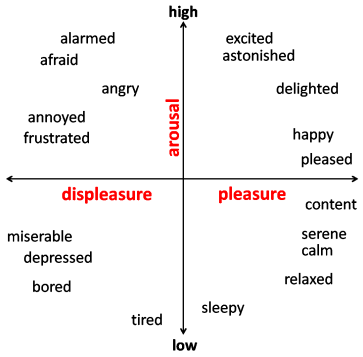

Fig. 1. Dimensional *arousal-valence* chart of human emotions.

tain facial expressions and action units [17]. This increased demand for automatic pain detection systems in health care and rehabilitation programs motivated researchers to collect benchmark datasets and develop approaches for pain detection and pain intensity estimation from facial expressions [18]. In addition to facial expressions, the body language has been recognized to convey a valuable part of emotions [19]. Several studies from different domains agreed that combining the face and body expressions can improve the recognition of emotional states [20], [21], [22]. However, exploiting the body language in conjunction with facial expressions is a topic which has been rarely investigated in the literature of automatic emotion recognition. Finally, the urgency for spontaneous datasets is now clearly recognized. In fact, most of the current solutions for facial expression recognition from 3D dynamic data are evaluated in constrained scenarios, which include high-resolution posed datasets acquired with rigid settings [1]. Instead, the recognition of spontaneous facial expressions is a more challenging problem that recently attracted high interest [23], [24]. The effect of low-resolution and noisy acquisitions on expression recognition is a related aspect that remained almost unexplored in these studies, while it is actually of increasing relevance when moving from constrained to real scenarios. Based on the above considerations, in this work, we target the problem of online emotion detection from non-posed 3D dynamic data, which include the face and the upper part of the body. As related problem, we focus also on early pain detection from spontaneous high-resolution 3D dynamic data. Both the scenarios are adapted to the same framework, which is based on subspace trajectory analysis.

The rest of the paper is organized as follows: Works related to the proposed approach are summarized in Sect. 2; In Sect. 3, we outline our main ideas and contributions; In Sect. 4, the proposed representation of video sequences as trajectories on a Grassmann manifold is presented; In Sect. 5, the 3D video representation is adapted to an early event-detector framework; The pain detection from 4D data is presented in Sect. 6; Experimental analysis of the proposed solution in reported in Sect. 7. Finally, conclusions and future work are discussed in Sect. 8.

2 Related Work

In the following, we consider works addressing two distinct but related topics, which are relevant to our proposal: Spontaneous emotion detection and classification from 3D videos; and pain detection from dynamic data.

Spontaneous emotion detection and classification: Facial expression classification and emotional states detection focused for long time on acted facial expressions due to the difficulty of collecting and annotating spontaneous and natural facial expression databases. Recently, more attention has been paid to the analysis of spontaneous facial expression and emotion detection as shown by the release of several databases [25], [26], [27]. Also, some databases appeared, which try to bring spontaneous facial expressions from 2D to 3D such as Sherin et al. [28], Zhang et al. [29], and Mahmoud et al. [30]. In the last few years, several works addressed the problem of spontaneous facial expressions analysis in the continuous space. Cruz et al. [31] proposed a bio-inspired approach for spontaneous facial emotion analysis with evaluation on the continuous space. This approach payed more attention to the parts of the scene with the highest dynamics by unfixed video down-sampling rate. Zeng et al. [32] proposed a one-class classification problem to distinguish between emotional facial expressions and non-emotional ones, with validation on continuous Adult Attachment Interview (AAI) database. Hupont et al. [33] presented a framework to find mapping between the six expressions categorization and the 2D continuous representation by the confidence value of the basic expression in the Ekman representation. The methods above limited their use to the facial expression modality only. Other methods tried to incorporate further modalities beside the facial expression. For example, Metallinou et al. [34] addressed the challenge of tracking continuous levels of a participant's activation, valence and dominance during the course of affective dyadic interactions using bodily and vocal information. Wöllmer et al. [35] proposed a framework for recognizing spontaneous emotions in continuous space of four dimension (arousal, expectancy, power and valence) from audiovisual data using Long Short Term Memory LSTM-modeling. The facial expression data are fused with the EEG signals for the generation of affective tags on the *arousal-valence* space in the work of Soleymani et al. [36]. Gaus et al. [37], estimated the arousal, valence and dominance of the affect from audiovisual information using wavelet analysis and Partial Least Squares regression. Mou et al. [38] proposed to estimate the arousal-valence value from an image of a group of people from the face and body visual data. This short summary evidences the importance of this new trend in affect analysis. We also note most of these works used 2D data, while incorporating the upper part of the body with the face is not investigated yet in 3D videos. A more detailed survey about emotion analysis in continuous space can be found in Gunes et al. [16].

Physical pain detection in videos: Physical pain detection and assessment from facial expressions attracted attention recently due to its important applications in health care systems, clinical treatment [39], [40] and the ability of recognizing pain affect from facial cues [41], [42]. Lucey et al. [43] presented a facial video database (known as UNBC-McMaster Shoulder Pain Expression archive) for people suffering from shoulder pain, with action unit coding on the frame level of the video. The same authors extended the work by proposing an Active Appearance Model (AAM)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

system that can detect the frame with pain expression out of others in 2D texture videos [44]. A full automatic pain intensity estimation approach from 2D image sequences is presented by Kaltwang et al. [45] on the same database. Khan et al. [46] presented a new facial descriptor called pyramid local binary pattern (PLBP), with application on pain detection. Their approach gives near real-time performances, with high recognition rate. Unlike previously mentioned works, Sikka et al. [47] proposed sequence level spatial-temporal descriptor instead of frame level to exploit the advantage of temporal information in the 2D video in combination with bag-of-words framework. This approach gives better results on McMaster Shoulder Pain database approving the positive effect of temporal information on recognizing pain. Since the works listed above are based on 2D images, they are affected by pose and illumination variations, which can be solved by moving to 3D imaging systems. Following other facial computer vision problems, pain recognition may be considered in 3D facial databases. In BP4D-Spontaneous 3D dynamic database proposed by Zhang et al. [29], there is one task of spontaneous physical pain experience for 41 subjects. Zhang et al. [48] proposed a pain related action units detection on BP4D database using binary edge feature representation. This approach exploits the available temporal information alongside the 3D facial scans as well their robustness against pose variation. A more comprehensive survey on pain detection from facial expressions can be found in Aung et al. [18]. As final comment, we can observe that, similarly to facial expressions, facial analysis for pain detection is following the shift from 2D to 3D, with early detection of pain out of 3D videos not investigated yet.

3 Methodology and Contributions

In this work, we propose an online emotion detection approach, capable of early detecting spontaneous human emotions from dynamic sequences of 3D/depth frames. The proposed framework is evaluated in two challenging problems: *(a)* Early detection of spontaneous emotional states from depth sequences of the upper part of the body acquired with a low-resolution sensor so that the emotions depend jointly on the dynamics of facial expressions and upper body; *(b)* Early detection of spontaneous physical pain from dynamic sequences of 3D high resolution facial sequences. In doing so, we introduce a new representation of human space-time 3D/depth data and the related processing tools. In fact, several inherent challenges arise in the analysis of 3D/depth videos, the most relevant one being related to the non-linearity of space-time data due to non-rigid face deformations or body gestures. In addition to rigid and non-rigid transformations, other challenges derive from missing or noisy data originated by auto-occlusions of the body and the acquisition itself, respectively. In the literature, solving these issues requires pose normalization as well as temporal registration along the depth-video, which are time consuming when processing dense data [49].

In our proposed approach, we account for the non-linearity of the data and related transformations as follows: First, we assume linearity in a local (short-time) interval, by grouping the depth frames into subsequences of predefined length and regarding each group as a linear subspace (i.e., span of an orthonormal basis, represented by a matrix), which gives rise to an element on a Grassmann manifold [50], [51]; Then, we generalize it to longer videos using curves (i.e., non-linear representation) on the underlying curved manifold. This manifold-mapping allows faithfully representing the original depth and 3D video data in a computationally economical way, showing robustness also to noisy and missing data [52]. This latter aspect makes the proposed representation suitable for processing and analyzing videos acquired with depth-consumer cameras, which suffer from low-accuracy, noisy depth measurements, and incomplete data. Finally, using a Structured Output SVM (SO-SVM) based on sequential analysis of Euclidean spatio-temporal features, our framework can perform early affect state detection online. Figure 2 summarizes the idea of mapping short-time depth video subsequences to a Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$, where k is the dimension of subspaces, and n the ambient space dimension. The position of points corresponding to successive subsequences captures the temporal evolution (i.e., dynamics) of the face or the body in 3D videos, shown as a trajectory on the manifold.

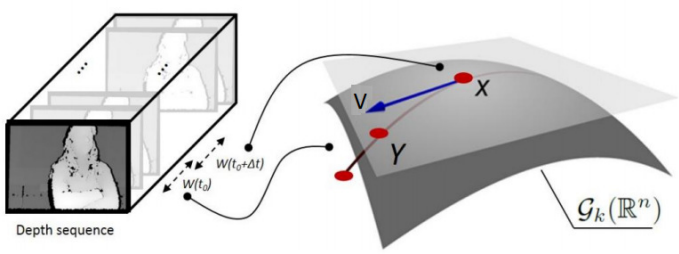

Fig. 2. Representation of dynamic depth data as a trajectory on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$.

In summary, the main contributions of this work are:

- A novel representation based on trajectories on Grassmann manifold, which can model 3D/depth sequences and inherent human motion (deformations, gestures, etc.) of non-linear nature;
- A new space-time feature representation, which captures human movements (both deformations and pose) suitable for analyzing dynamic facial or body data;
- An adaptation of the SO-SVM early event detector [53] for sequential analysis of Grassmann trajectories;
- Jointly consider the upper body movement and the face to detect complex spontaneous emotions from depth videos acquired with a cost-effective Kinect camera;
- Early detection of spontaneous physical pain from high-resolution dynamic sequences of 3D faces.

4 Mathematical Framework

In this work, we define a dynamic subspace representation approach capable of modeling the spatio-temporal information of both high-resolution 3D or low-resolution depth sequences. In the following, we refer to the general case of depth images, since high-resolution 3D scans are cast to this case by mapping to depth frames for subsequent processing. We consider a continuous dynamic flow of depth frames, where a region of interest (e.g., the face alone or

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

the entire upper part of the body including the head) can be detected. If a local (short-time) interval is considered (i.e., a temporal window of few frames), we can assume the region of interest preserves a constant size across the frames of the window. Let $n = w \times h$ be the size (i.e., number of depth pixels) of the region of interest, with width w and height h, for the generic frame at instant t. This region can be reshaped and regarded as a vector $f_t \in \mathbb{R}^n$. Based on this temporal locality, the depth frames can be grouped into subsequences of predefined window length ω and cast to a matrix representation, where the individual frames are columns of the matrix. Accordingly, given the *window* size ω, a subsequence of frames starting at t_0 is modeled as a matrix $\mathbf{F} = [f_{t_0}, f_{t_0+1}, \ldots, f_{t_0+\omega-1}] \in \mathbb{R}^{n\times\omega}$. This representation retains all the depth data of a subsequence, thus it is likely to include redundant information. Applying Singular Value Decomposition (SVD) to the matrix $\mathbf{F}$ permits us to retain the important information spanned by the k-first singular vectors and compensate possible missing data in some frames [52]. These principle vectors constitute an orthonormal basis X that span the subspace $\mathcal{X}$. More formally, $\mathcal{X} = Span(X) = Span(\{v_1, \ldots, v_k\})$, where v_i is the ith singular vector and the $Span(.)$ of an orthonormal base is precisely the subspace defined by this base. In turn, $\mathcal{X}$ is an element on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$ [50], which can be intuitively defined as the set of k-dimensional linear subspaces of $\mathbb{R}^n$ (see Fig. 2). The depth frames subsequences compact representation as elements on the Grassmann manifold is the basic operation at the core of our proposed framework. The intrinsic advantage of this representation is its capability of naturally handling pose variations and missing parts, jointly with changes due to face and body deformations.

More in detail, the Grassmann manifold can be derived from the *Stiefel* manifold, which provides a representation for the matrix subspace. So, in the following, we will refer to the matrix representation of subspaces (i.e., elements of the Stiefel manifold $\mathcal{V}_k(\mathbb{R}^n)$), since the extension to $\mathcal{G}_k(\mathbb{R}^n)$ is straightforward [50]. Formally, let us consider the set of $n\times k$ tall-skinny orthonormal matrices (of linearly independent vectors) in $\mathbb{R}^n$. The space of such matrices endowed with a Riemannian structure is known as *Stiefel manifold* $\mathcal{V}_k(\mathbb{R}^n)$ and defined as follows:

$$\mathcal{V}_k(\mathbb{R}^n) \triangleq \{X \in \mathbb{R}^{n\times k} \colon X^T X = I_k\} \ . \tag{1}$$

Under this definition, a k-dimensional subspace of $\mathbb{R}^n$ is spanned by all rotations of the element $X \in \mathcal{V}_k(\mathbb{R}^n)$ under the orthogonal group $\mathbb{O}(k)$. Grouping together all possible rotations of every point on Stiefel manifold $\mathcal{V}_k(\mathbb{R}^n)$ gives rise to the *Grassmann manifold* denoted by $\mathcal{G}_k(\mathbb{R}^n)$. Stated differently, the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$ is defined as the quotient space of the Stiefel manifold as follows:

$$\mathcal{G}_k(\mathbb{R}^n) \triangleq \mathcal{V}_k(\mathbb{R}^n)/\mathbb{O}(k) \ , \tag{2}$$

where two points X, Y on $\mathcal{V}_k(\mathbb{R}^n)$ are from the same equivalence class (i.e., $X \sim Y$) if their columns span the same k-dimensional subspace, that is $Span(X) = Span(Y)$. Also, an orbit of $\mathcal{V}_k(\mathbb{R}^n)$ under the group action $\mathbb{O}(k)$ represents the same point on $\mathcal{G}_k(\mathbb{R}^n)$.

With the proposed representation, a depth video is partitioned into adjacent subsequences of size ω, and thus into a set of points on the manifold. Then, it is quite natural to look to the temporal sequence of points as describing a *trajectory* on the manifold (see points X and Y in Fig. 2). The analysis of motion information of depth videos is thus mapped to the problem of analyzing smooth trajectories on the manifold using mathematical tools which account for the non-linear geometry of the manifold. In this context, Riemannian manifold representations revealed great success in different computer vision problems [54].

A first and straightforward approach for capturing the temporal evolution along a trajectory is to compute the *geodesic distance* along it. Given arbitrary elements $X, Y \in \mathcal{V}_k(\mathbb{R}^n)$ and $\mathcal{X} = Span(X)$, $\mathcal{Y} = Span(Y) \in \mathcal{G}_k(\mathbb{R}^n)$, Golub and Loan [55] introduced an intuitive and computationally efficient way of defining the distance between two linear subspaces using the *principal angles*. In fact, there is a set of principal angles $\Theta = [\theta_1, \ldots, \theta_k]$ between $\mathcal{X}$ and $\mathcal{Y}$, defined as follows:

$$\theta_i = \cos^{-1}\left(\max_{u_i\in\mathcal{X}} \max_{v_i\in\mathcal{Y}} \langle u_i^t, v_i\rangle\right) \ . \tag{3}$$

In this equation, u and v are the vectors of the basis spanning, respectively, the subspaces $\mathcal{X}$ and $\mathcal{Y}$, subject to the additional constraints $\langle u^t, u\rangle = \langle v^t, v\rangle = 1$, and $\langle u^t, v\rangle = \langle v^t, u\rangle = 0$, where $\langle ., .\rangle$ denotes the *inner product* in $\mathbb{R}^n$. Figure 3 illustrates two subspaces $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^n$, and the principal angles $\Theta = [\theta_1, \ldots, \theta_k]$ between them.

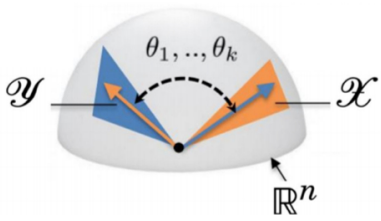

Fig. 3. Principal angles $\theta_1, \ldots, \theta_k$ between two k-dimensional subspaces $\mathcal{X} = Span(X)$ and $\mathcal{Y} = Span(Y) \in \mathbb{R}^n$.

Based on the definition of the principal angles, the geodesic distance $d_{\mathcal{G}}$ between $\mathcal{X}$ and $\mathcal{Y}$ can be defined by:

$$d^2_{\mathcal{G}}(\mathcal{X}, \mathcal{Y}) = \sum_i \theta_i^2 \ . \tag{4}$$

In the case two elements are given on the Stiefel manifold, that is $X, Y \in \mathcal{V}_k(\mathbb{R}^n)$, the distance $d_{\mathcal{V}}$ can be defined more appropriately by the standard *Chordal* distance:

$$d_{\mathcal{V}}(X, Y) = \|X - Y\|_F \ , \tag{5}$$

where $\|\ .\ \|_F$ is the *Frobenius* norm $\|X\|_F = \sqrt{tr(XX^t)}$. The first idea that we explore here is the use of the *geodesic distance* computed along a trajectory to (globally) capture the dynamics of the human body. We shall then use this temporal signature to learn an online detector (see Sect. 5). The second idea, presented in Sect. 6, is to use initial velocity vectors computed along the trajectory (see v in Fig. 16), instead of the geodesic distances (i.e., norms of the velocity vectors $\|v\|$), once transported to a tangent plane attached to $\mathcal{G}_k(\mathbb{R}^n)$ in one reference point. This mathematical framework provides us the background to derive a precise representation of parameterized trajectories on the manifold.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

5 Emotion Detection from Depth-Bodies

The basic framework defined above is adapted in the following to the case of low-resolution depth videos of the upper part of the body (face, neck, shoulders and arms/hands), which are acquired with a depth camera (Kinect).

In a first step, the upper part of the body is segmented from the background in each depth frame of the observed video. Then, the depth sequence of the cropped upper body is divided into successive short-time subsequences, based on a temporal window size ω. In each subsequence of size ω, the depth information of the frame is reshaped as a vector of size n, which is then arranged to build a matrix X of size $n \times \omega$. Applying k-SVD to X gives us $X = U\Sigma V^T$. The subspace spanned by the k-columns of the matrix U is retained to represent the original subsequence. As a result, every complete depth video is divided into m short subsequences of ω frames. Thus, each video is mapped to m k-linear subspaces, which are points on a Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$. This representation by trajectories on Grassmann manifold allows us to reduce the effect of the noise of the acquired depth data, and constitutes an efficient way to sequentially analyze the observed video stream, and extract relevant space-time features for online emotion detection.

More specifically, trajectories on the Grassmann manifold can be analyzed by considering the evolution of their instantaneous speed. Given an observed portion of the trajectory in the time interval $[0, t]$, the instantaneous speed can be computed as the distance between neighboring points $\mathcal{X}^t$ and $\mathcal{X}^{t+\delta}$ along the trajectory, with δ assuming integer values, $\delta = \{1, 2, 3, \dots\}$. The length of the shortest path between successive subspaces along the trajectory is computed using the geodesic distance in Eq. (4), and the standard chordal distance given in Eq. (5), for the Grassmann and Stiefel manifold, respectively. These distances are accounted in an array characterizing the temporal evolution along the trajectory, that we call *Geometric Motion History* (GMH). The main idea behind this representation is that similar emotional states will be conveyed by similar movements and facial expressions, which leads to similar GMH vectors and vice versa.

5.1 Structured Output learning from sequential data

Early detection from sequential data aims to find the correct classifier capable of providing a decision from both partial and complete events. This should permit recognition of the emotion of interest, while receiving sequential data, and also provide the initial and ending time of the event. To this end, we adopted the Structured-Output SVM (SO-SVM) [53] framework, motivated by some interesting aspects of this classifier: *(i)* it can be trained on all partial segments and the complete one at the same time; *(ii)* it allows us to model the correlation between the extracted features and the duration of the emotion; *(iii)* no previous knowledge is required about the structure of the emotion; *(iv)* it can give better performance than other algorithms in sequence-based applications [56]. Further details about the structured output learning framework can be found in Hoai et al. [53]. Assume a set of GMH feature vectors are computed, each including an emotion of interest. The start and end time of the emotion are also annotated by a pair of values $[s^i, e^i]$. At any time t^i comprised between the start and end of the emotion $s^i \leq t^i \leq e^i$, all partial emotions sub-segments obtained between $[s^i, t^i]$ will be used to train the SO-SVM detector, since all these different size sub-segments represent positive samples. All the other parts of the GMH are considered as negative samples. The expected performance from SO-SVM in the testing stage is to fire the detection of the emotion of interest as soon as possible (after it starts and before it ends).

Algorithm 1 summarizes the steps of our proposed method for early emotion detection from depth bodies.

Algorithm 1 – Online emotion detection from depth-bodies

Require: Set $\mathbb{S} = \{S^i_{m_i}\}_{i=1}^{M}$ of depth body videos $S^i_{m_i}$, each with m_i frames; The segmenting window size ω

Initialization

for $i \leftarrow 1$ to M **do**

- $\hat{S}^i \leftarrow S^i$ *// depth preprocessing*
- $X_i\{\mathcal{X}^i_1, \mathcal{X}^i_2, \dots, \mathcal{X}^i_N\} \leftarrow \hat{S}^i$ *// video subsequences*
- $\mathcal{T}_i\{1, \dots, N\} \leftarrow$k$-SVD(X_i\{1, \dots, N\})$ *// trajectory building*
- $GMH_i \leftarrow distance(\mathcal{T}_i)$ *// compute distances*

end for

Processing

- $D\{i\} = [GMH_i]$ *// GMH of the event*
- $Label\{i\} = [s,\ e]$ *// event boundaries*
- Model = SO-SVM($D_{tr}, Label_{tr}$) *// SO-SVM training*
- y* = SO-SVM(D_{tst},Model) *// SO-SVM testing*

Ensure: y* = [s*, e*] *// detected boundaries of the event*

6 Physical Pain Detection from 4D-Faces

In the following, the framework presented in Sect. 4 is adapted to the case of spontaneous physical pain detection from dynamic sequences of high resolution 3D scans. Two different representations of facial data are used here: the *baseline* method, which uses the 3D facial landmarks available in the video; and the method based on depth frames obtained from the sequence of high-resolution 3D face scans.

As an example, Figure 4 shows a textured 3D pain face with its landmarks and depth image. This scan is taken from the BP4D-spontaneous expression dataset [29].

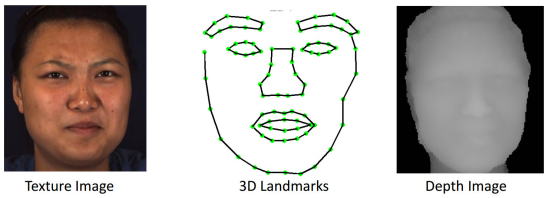

Fig. 4. From left to right: color image; 3D landmarks; and depth image.

6.1 3D landmark-based Grassmann trajectories

In this solution, we start from a sequence of high-resolution 3D face scans, each of which is labeled with N facial landmarks. The 3D coordinates (x, y, z) of the facial landmarks are considered as descriptor of the 3D facial scan, so that each frame is represented by a vector in $\mathbb{R}^{3\times N}$. Starting from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

this representation, and following the same steps of Sect. 5, we obtained a trajectory $\mathcal{T}$ of subspaces on a Grassmann manifold $\mathcal{G}_k(\mathbb{R}^{3\times N})$ for every 3D dynamic pain sequence. The motion information is then captured by computing the geodesic distance between successive subspaces by step δ to build the GMH of this video. This solution uses local and sparse information of the 3D shape of the face, and will serve as baseline to compare with the dense 3D shape representation using depth images.

6.2 Depth-based Grassmann trajectories

In this case, a depth image of the face is obtained from each high-resolution 3D scan after preprocessing and cropping of the facial area. Then, every subsequence of ω depth frames is modeled as a k-dimensional subspace in $\mathbb{R}^n$, being n the depth image size after vectorization. This permits us to build a time-parameterized trajectory $\mathcal{T}(t)$ of subspaces on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$, similarly to the case of Sect. 5. In this scenario, in addition to build the GMH by computing the geodesic distances between successive subspaces, like in the landmarks representation method, we propose a more efficient representation of the facial dynamic, called *Local Deformation Histogram* (LDH) descriptor, which is based on the concept of *vector field* of transported velocity vectors of trajectories on the manifold.

The speed along trajectories on the manifold (either Stiefel or Grassmann) allows us to quantify sequentially the motion amplitude and the temporal dynamics from depth data. However, to fully characterize the motion information the *field of velocity vectors* can be considered, instead of the distances used in Sect. 5, along the trajectories on the manifold. The main issue when using the velocity vector field along a trajectory is that they belong to different tangent planes, which makes difficult the use of a learning model on them. One intuitive solution will be to translate all the vectors of the field to the same tangent plane defined on a reference point, element of the Grassmann manifold, as described later. We will follow this solution in our approach. This computation requires the concept of *tangent space* to the manifold, *exponential map*, and *parallel translation*, which have been first defined in Rentmeesters et al. [57] and Shrivastava et al. [58]. For convenience, we first recall these definitions:

– **Tangent Space:** The tangent space at any X matrix representation of a subspace $\mathcal{X} \in \mathcal{G}_k(\mathbb{R}^n)$ is defined on the manifold as:

$$T_X(\mathcal{V}_k(\mathbb{R}^n)) = \{V \in \mathbb{R}^{nxk} \mid V^T X + X^T V = 0\} \,. \quad (6)$$

That is, $V^T X$ is a *skew-symmetric* $k \times k$ matrix.

– **Exponential Map:** It is a function that maps a point from the tangent space T defined at a point $X \in \mathcal{V}_k(\mathbb{R}^n)$ into the $\mathcal{V}_k(\mathbb{R}^n)$:

$$exp_X : T_X(\mathcal{V}_k(\mathbb{R}^n) \rightarrow \mathcal{V}_k \,. \quad (7)$$

The best way to present this function is to start from the geodesic connection between two points on the manifold, that is the shortest path connecting them. The derivative of this geodesic path at $t = 0$ is the initial tangent vector V as $\dot{\gamma} = V$, and the corresponding exponential map is $exp_X(V) = \gamma(1)$. This exponential map is defined only locally, and it can be computed by:

$$exp_X(V) = (XW \cos\Sigma + U \sin\Sigma)W^T \,, \quad (8)$$

where $V = U\Sigma W^T$ is the Singular Value Decomposition (SVD) of V.

– **Parallel Translation:** The parallel translation of a tangent vector T to $\mathcal{V}_k(\mathbb{R}^n)$ along a geodesic connecting $X = \gamma(0)$ and $Y = \gamma(1)$ denoted by $T_{X\rightarrow Y}$ is given by:

$$T_{X\rightarrow Y} = (-XW \sin\Sigma + \cos\Sigma)U^t T + (I - UU^t)T \,, \quad (9)$$

where $V = U\Sigma W^t$ is the compact SVD of the tangent vector V, such that $exp_X(V) = Y$ as defined in Eq. (8).

Based on these definitions, we start by computing the velocity vectors between neighboring subspaces along the trajectory (vector V in Fig. 2). However, these vectors belong to different tangent spaces. For example, the velocity vector between $\mathcal{X}(t)$ and $\mathcal{X}(t+\delta)$ is V, which lies on the tangent space $T_{\mathcal{X}}(\mathcal{G}_k(\mathbb{R}^n))$ defined at X, and so on. Since it is important to obtain a common (unique) vector field along the trajectory, one possibility is to translate the velocity vectors to one fixed tangent space to have them in the same vector space. In our solution, we translate all velocity vectors to the tangent space defined at the element of the manifold spanned by the identity matrix $I_k \in \mathbb{R}^{k\times k}$:

$$\mathcal{I} = Span\left(\begin{bmatrix} I_k \\ 0 \end{bmatrix}\right) \,, \quad (10)$$

where $\mathcal{I}$ is the chosen reference subspace on the Grassmann manifold $\mathcal{G}_k(\mathbb{R}^n)$ spanned by the identity matrix I_k. We chose $\mathcal{I}$ as reference for the parallel translation because it is well defined for all subspaces in the training and testing trajectories. Formally, after computing the velocity vector V between neighboring points on the trajectory, $\mathcal{X}^t$ and $\mathcal{X}^{t+\delta}$, with $V \in T_{\mathcal{X}}(\mathcal{G}_k(\mathbb{R}^n))$, we use the *parallel transportation* of Eq. (9) to translate it to the identity tangent space $T_I(\mathcal{G}_K(\mathbb{R}^n))$. Repeating this operation for all the velocity vectors along the trajectory results in an equivalent representation in one vector space (the tangent space attached to the identity element). Hence, the obtained transported velocity vector field reflects the motion of the face. According to Eq. (6), the velocity vector V_i is written as a matrix of size $n \times k$. Taking the k-first columns of this matrix V_i as vectors of size n and reshaping them to the original dimension of the face depth image $\widehat{m} \times \widehat{n}$ gives rise to k-first components. Visualizing the values of these components as 2D color mapped images shows clearly the temporal deformation with respect to spatial location in the original depth image. The first component of the velocity vector contains informative motion data, while the rest contains noise and redundant data. Then, rather than using the Grassmann distance that quantifies the speed along the trajectory, we propose to exploit the first component of the velocity vector between two subspaces. This new representation for the temporal evolution on the trajectory carries information not only about the speed (intensity) of the deformation, but also about where in the face and in which direction the deformation occurs. This is illustrated in Fig. 5 for a depth sequence of *physical pain*. In this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

Figure, a color between green-to-red indicates deformations in the forward direction of the face, while the green-to-blue color means that the deformation occurs in the backward direction. The static part of the face through the time (green color) is also identified and discarded.

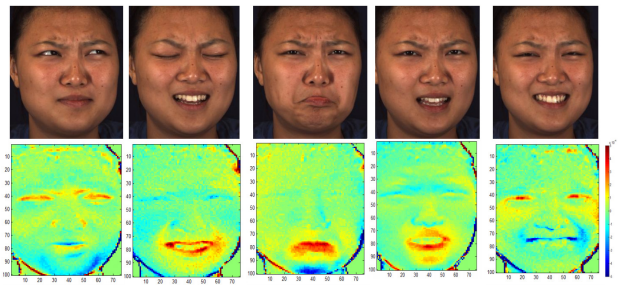

Fig. 5. Visual illustration of computed velocity vectors between subspaces (bottom) with the corresponding 2D texture images (top). The colors show the deformation areas and their direction: green colors mean absence of deformations; from green-to-red forward deformations; and from green-to-blue backward deformations.

In a final step, the matrix is divided into blocks, thus permitting us to localize where the deformation happens in the face, and a dual value (positive/negative) histogram is computed for each block. This histogram provides us a quantitative measure about the intensity of the deformation of the facial region associated to the block in the two directions. The concatenation of the histograms of all blocks provides what we call the *Local Deformation Histogram* (LDH) from the velocity vector. The LDH vectors between each two subspaces on the same trajectory are concatenated to build a general LDH descriptor of the trajectory $\mathcal{T}$ on the Grassmann manifold. Figure 6 illustrates these steps.

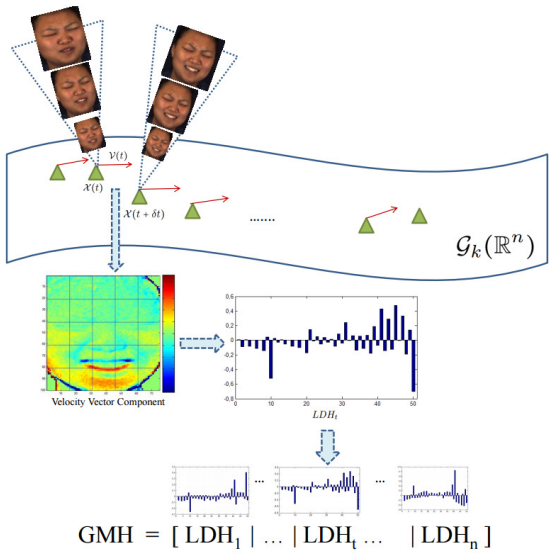

Fig. 6. Illustration of LDH computation from the velocity vectors (red arrows) between subspaces (green triangles) of the same trajectory.

The start and the end time of the physical pain event is decided depending on certain annotated facial action units combination (this aspect will be discussed in more detail in Sect. 7). The SO-SVM approach presented in Sect. 5.1 is used to detect the pain feeling as early as possible from the GMH features extracted from the landmarks and depth representation. Algorithm 2 summarizes the approach.

Algorithm 2 – Physical pain detection from 4D-faces

Require: Set $\mathbb{S} = \{S^i_{m_i}\}_{i=1}^{M}$ of 4D facial scans $S^i_{m_i}$ each with m_i frames; window size ω; $Labels\{L^i\}_{i=1}^{M}$, where $L^i[s,e]$ indicates the start and the end of pain affect in S^i

Initialization

for $i \leftarrow 1$ to M **do**

$\hat{S}^i \leftarrow S^i$ *// 3D preprocessing and depth generation*

$X_i\{\mathcal{X}^i_1, \mathcal{X}^i_2, \ldots, \mathcal{X}^i_N\} \leftarrow \hat{S}^i$ *// video subsequences*

$\mathcal{T}_i\{1,..,N\} \leftarrow k\text{-}SVD(X_i\{1,\ldots,N\})$ *// trajectory building*

$\mathcal{V}_i \leftarrow Velocity(\mathcal{T}_i)$ *// velocity vectors between subspaces*

$\mathcal{V}_i^T \leftarrow Transport(\mathcal{V}_i)$ *// tranportation to one tangent space*

$LDH_i\{1,..,N\} = LDH(\mathcal{V}_i^T)$ *// LDH from velocity vectors*

$LDH_i \leftarrow [LDH_i(1), LDH_i(2), \ldots, LDH_i(N)]$ *// GMH*

end for

Processing

Model = SOSVM($LDH_{tr}, Labels_{tr}$) *// SO-SVM training*

y* = SOSVM(LDH_{ts}, Model) *// SO-SVM testing*

Ensure: y* = [s*, e*] *// boundaries of the detected pain affect*

7 Experiments and Evaluation

To validate the proposed framework, we have conducted experiments on two different datasets. The first dataset [30] includes depth-videos of the upper part of the body, when spontaneous emotions or complex mental states, like happiness, thinking, etc, are exhibited. We apply our framework on this dataset in order to obtain early detection of spontaneous emotional states. The second dataset [29] includes high resolution 3D videos of faces showing also spontaneous affects, like happiness, sadness, physical pain, etc. On this database, our experiments focus on early detection of spontaneous physical pain using different representations. Performances are measured in terms of accuracy and timeliness using the following evaluation criteria:

- **Area under the ROC (AUC) curve:** A ROC curve is created by plotting *True Positive Rate* (TPR) vs. *False Positive Rate* (FPR) at varying threshold. The AUC curve gives the overall performance of the binary classifier to discriminate between positive and negative samples;
- **AMOC curve:** The *Activity Monitoring Operating Characteristic* curve is generally used to evaluate the timeliness of any event surveillance system. It gives an indicator of how much the detection of the event is fast, by reporting the *Normalized Time to Detection* (NTtoD) as a function of False Positive Rate (FPR). In particular, NTtoD is defined as the fraction of the event occurred at one time instance. For an event starting at s and ending at e in a time series, if the detector fires the event at time t where $s < t < e$, the NTtoD is given by:

$$NTtoD = \frac{t-s+1}{e-s+1}\,. \tag{11}$$

- **F1-score curve (or F-measure):** defined as the weighted harmonic mean of its *precision* and *recall*. It is high only when both recall and precision are high.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

7.1 Cam3D Kinect database

In the Cam3D Kinect database, Mahmoud et al. [30] collected a set of 108 audio/videos of natural complex mental states of 7 subjects. Each video is acquired with the Kinect camera, including both the appearance (RGB) and depth (D) information. The data capture natural facial expressions and the accompanying hand gestures, which are more realistic and more complex than the basic six facial expressions. Figure 7 shows example frames for four emotional states.

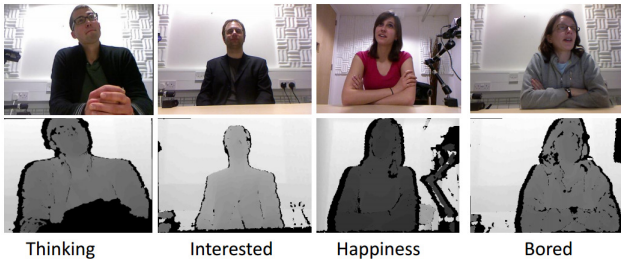

Fig. 7. Cam3D Kinect database: Examples of depth frames with their corresponding 2D texture image of different emotional states.

Table 1 summarizes the number of available videos for each emotional state. These videos provide a sampling of the dimensional description chart of emotions in Fig. 1. Several categories include few videos (i.e., less than 8 videos are present in 9 out of the 12 emotion categories, with 5 categories including just 1 or 2 videos), thus precluding their use in detection experiments. This motivated us to consider the following two experimental scenarios: *Happiness* vs. *others*; and *Thinking/Unsure* vs. *others*. Compared to the *arousal-valence* chart of Fig. 1, the first scenario tests the detection of an emotion located in the *high-arousal / pleasure* quadrant (positive emotion); the second one refers to an emotion in the *low-arousal / displeasure* sector (negative emotion). We grouped *Thinking/Unsure* together since they belong to the same group of affects called *Cognitive group* as categorized in Mahmoud et al. [30].

TABLE 1
Number of available depth videos for each emotional state in Cam3D

Emotional/Mental State	# of depth videos
Agreeing	4
Bored	3
Disagreeing	2
Disgusted	1
Excited	1
Happy	**26**
Interested	7
Neutral	2
Sad	1
Surprised	5
Thinking	**22**
Unsure	**32**

7.2 Emotional state detection

We applied the speed along trajectories (GMH feature) on the manifold (see Algorithm 1) to detect emotional states from two different regions of the dimensional *arousal-valence* emotion chart of Fig. 1: *Happiness* vs. *others* and *Thinking/Unsure* vs. *others*. In both experiments, the videos of the emotion of interest and the videos of the other emotions are divided equally into two halves, one used for training and one for testing in a subject-independent manner. Then, the GMH feature is computed by dividing each video into subsequences of size $\omega = 20$ and subspace dimension $k = 5$ (this setting has been chosen empirically). Then, the GMH of the emotion of interest is concatenated with the GMH computed for two videos of different emotional states (see Fig. 8). Selecting these videos randomly for each concatenation, permitted us to obtain more training and testing data. We derive a total of 100 GMH for training, and the same number for testing. For each generated sequence, the start and the end point of the emotion of interest is known.

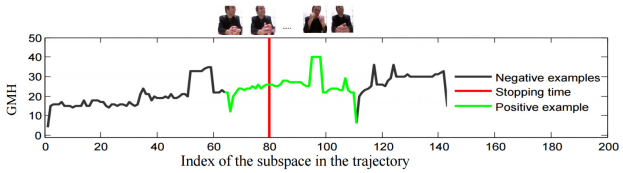

Fig. 8. An example of the GMH feature vector on Grassmann manifold. The GMH for the emotion of interest (*Happiness*) is in the middle (green) between two other different emotions. The early online detection occurs in correspondence of the red line.

In a first experiment, we compare the sequential analysis of trajectories using the proposed GMH feature computed for the Grassmann and Stiefel manifolds. For the *Happiness* vs. *others* case, the top of Fig. 9 shows the ROC and the AMOC curves obtained. From the ROC curves related to the Grassmann, it can be observed that when the FPR is around 20% the TPR reaches 70% for *Happiness* detection. This accuracy decreases significantly (around 50%) at FAR=10%. Comparing the analysis of the trajectories along the Stiefel (dashed curves) and the Grassmann manifold (continuous curves), it clearly emerges the sequential analysis performed on Grassmann manifold outperforms the analysis on Stiefel manifold. The area under the ROC curves (AUC) is 0.73 and 0.84 on Stiefel and Grassmann, respectively. The same conclusion can be obtained by comparing Stiefel and Grassmann manifolds for the *Thinking/Unsure* emotional state in the bottom of Fig. 9.

This demonstrates the consistency of the subspace based representation $\mathcal{Y} = Span(Y)$ and the associated metric $d_{\mathcal{G}}$ over the matrix representation. This is mainly due to the invariance of the subspace representation to rotations $\mathbb{O}(k)$ as $\mathcal{G}$ is a quotient space of $\mathcal{V}$ under the group action of $\mathbb{O}(k)$. The plots on the right of Fig. 9 show the evolution of the system latency (the fraction of video needed to make the binary decision) against FPR. For example, the detector achieves 20% of FPR by analyzing 20% of the video segment. Also in this case, results reported for the Grassmann representation are better than results obtained for the Stiefel representation.

Comparing detection accuracy results for *Happiness* and *Thinking/Unsure* from Fig. 9, the *Thinking/Unsure* detection shows a performance decrease with respect to the *Happiness* detection. The area under the ROC curve (AUC) is 0.66 and 0.79 on Stiefel and Grassmann manifold, respectively, for *Thinking/Unsure*, while they are 0.73 and 0.84 for *Happiness*. These results confirm the advantage in using the Grassmann rather than the Stiefel representation. From the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

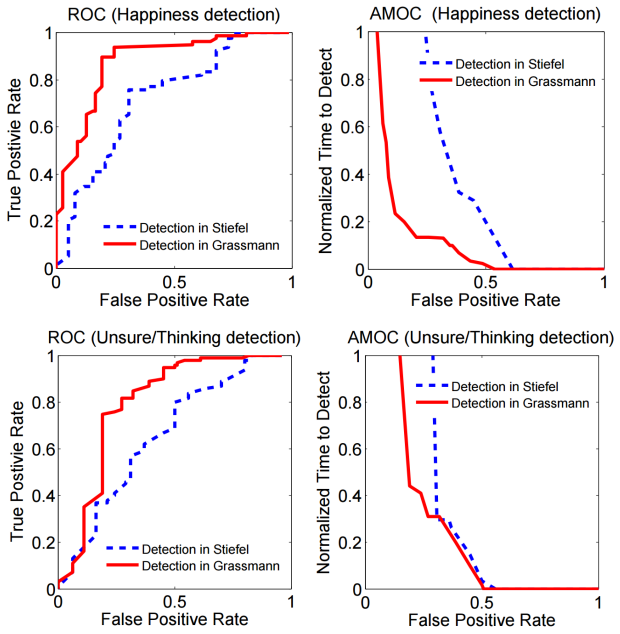

Fig. 9. ROC and AMOC curves for *Happiness* detection (top) and *Thinking/Unsure* detection (bottom) over Stiefel and Grassmann manifolds.

plot on the right of Fig. 9, it can be noted that about 20% of the negative samples are recognized to be element of this class, even if the videos are observed completely. This can be motivated by the "common" neutral behavior exhibited by humans when conveying other complex mental states (e.g., agreeing, bored, etc.). This was not the case for the *Happiness* detector, as the happiness is often accompanied by body and facial expressions.

To investigate the importance of using the upper part of the body (face, shoulders and hands) versus using only the face, we performed experiments with the previous protocol, but considering the upper body in the depth videos to construct the GMH on Grassmann manifold, instead of the cropped region of the face only. From Fig. 10, it is clear that the emotional state exhibited by the upper body is easier to detect than considering the facial region alone, when acquired using cost-effective cameras. In the *Happiness* experiment, the area under the ROC curve for the upper body and the face only are 0.84 and 0.68, respectively. Performing the same experiment for the *Thinking/Unsure* case, the area under the ROC curve is 0.79 and 0.63 for the upper body and the face only, respectively. This result is in agreement with studies like Stock et al. [20] and Meeren et al. [21], which encourage the use of the upper body with the face in automatic emotional state understanding.

Finally, we also investigated the relevance of the window size ω (number of frames used to embody the motion in the subspace). Empirically, we found the best window size for this application is $\omega = 20$. In Fig. 11, we compare $\omega = 20$ and $\omega = 5$ (red and blue curves, respectively) for the Grassmann manifold and *Happiness* emotion detection. The dimension of the subspace is $k = 5$ in both cases. In the first case, with $\omega = 20$, using five singular values permits us to keep 90% of the original information of the temporal window (we selected this value by empirical experiment); in the case

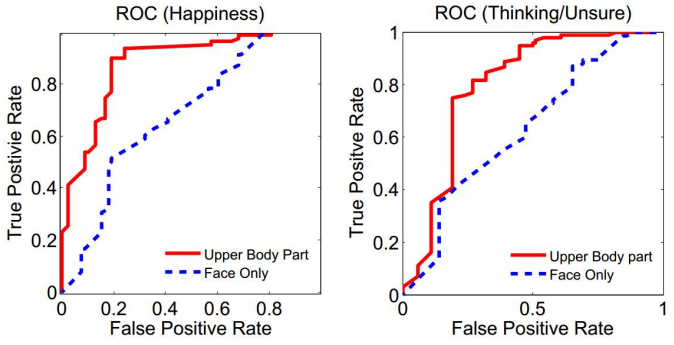

Fig. 10. ROC curves comparison for *Happiness* and *Thinking/Unsure* detection over the Grassmann manifold using the upper body, and the face only.

of $\omega = 5$, we keep 100% of the information as $k = \omega = 5$. The area under the ROC curve for $\omega = 5$ is 0.74, and 0.84 when $\omega = 20$. A small window size, such as $\omega = 5$, for these depth videos captured with temporal resolution of about 25fps leads to very slight differences between successive subspaces and, as a result, less representative GMH features. Results for more window size are reported in Table 2.

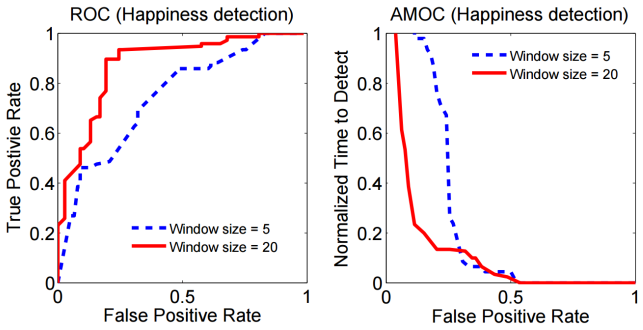

Fig. 11. ROC and AMOC curves for *Happiness* detection over the Grassmann manifold for two different window size.

TABLE 2
Area under ROC curves with varying window size ω

ω	**5**	**10**	**15**	**20**	**25**	**30**
AUC	0.74	0.78	0.81	0.84	0.82	0.80

In order to investigate the statistical significance of our proposed method, we repeated 100-times the previous experiments with the optimal parameters ($\omega = 20$ and $k = 5$). In each run, the negative examples before and after the positive examples (emotion of interest) are randomly selected. F1-score ($\pm$ standard deviation) is reported against the fraction of the video seen. Results are shown in Fig. 12, for the *Happiness* and *Thinking/Unsure* detectors (red and blue curves, respectively). For short fractions of the event seen, the two cases show similar behavior, while the *Happiness* result is clearly better than the *Thinking/Unsure* result when the fraction of the event increases.

Finally, we compared our proposed geometrical framework using depth channel, with 2D channel for happiness and thinking early detection from the upper part of the body. We divided the 2D videos into sub-sequences of size

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

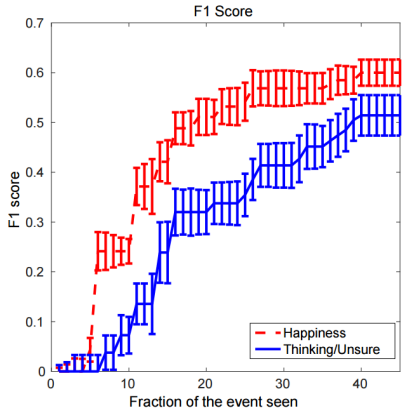

Fig. 12. Average F1-scores (with standard deviation) obtained for the *Happiness* emotion and the *Thinking/Unsure* affective states against the fraction of the event seen.

$\omega = 20$ as for the depth channel, and applied the LBP-TOP descriptor for dynamic 2D video analysis as presented in Zhao et al. [59]. From this feature, we extract the XY, XT and YT feature planes, where X and Y are the horizontal and vertical image axis, respectively, and T is the time. Every sub-sequence is represented by the resulting histogram. Concatenating all histograms of the subsequences gives rise to a feature vector presented to the same early detector, for comparison. From Fig. 13, we can see the AUC for *Happiness* and *Thinking/Unsure* detection is better using our geometric framework than using the 2D channel with the LBP-TOP feature (i.e., 0.84 and 0.79 using our method, compared to 0.72 and 0.69 using the LBP-TOP, for *Happiness* and *Thinking/Unsure*, respectively).

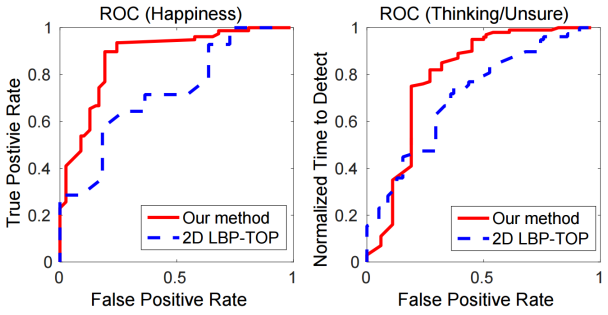

Fig. 13. ROC curves for *Happiness* and *Thinking/Unsure* early detection using our method on depth data and LBP-TOP on 2D data.

7.3 BP4D-Spontaneous facial expression database

Zhang et al. [29] proposed Binghamton-Pittsburgh 3D dynamic (4D) spontaneous facial expression database. This database includes 41 subjects acquired using Di4D dynamic face capturing system at 25fps. There are 8 different tasks for every subject corresponding to the following spontaneous expressions: *Happiness* or *Amusement*, *Sadness*, *Surprise*, *Embarrassment*, *Fear* or *Nervous*, *Physical pain*, *Anger* or *upset* and *Disgust*. This database provides the 3D model and the 2D still images for every video with metadata. Metadata include, for 2D texture images, the 46 landmarks annotation with the pose information and, for 3D models, 83 feature points (landmarks) annotation with the pose information given by the *pitch*, *yaw* and *roll* angles. Facial action units (FAUs) are provided for 20 seconds (about 500 frames) of every task. This AU annotation provides information about specific AUs activation in the frame and their intensity in the case of activation. Figure 14 depicts one 3D model with its corresponding 2D texture image for every task.

Fig. 14. BP4D Database: Examples of the eight different spontaneous expressions (tasks) included in the database.

7.4 Analyzing 4D-Faces for Physical Pain Detection

We applied the proposed geometric framework with the transported velocity vector field method, as explained in Sect. 6 to detect spontaneous physical pain from 3D dynamic facial videos. The spontaneous physical pain in the BP4D database is elicited by putting the participant's hand in ice water. The acquired 3D videos are quite long (their duration is about 20s), and it is known there is a pain emotion through the video, which constitutes our initial ground truth. To have accurate pain affect start and end points during the video as an emotion of interest, we use the FAUs provided annotation. Several studies have been conducted in psychology field to reveal the optimal AUs combination that can define the physical pain emotional state. Prkachin et al. [60] proposed a pain intensity scale equation (PSPI) considering certain AUs given by:

$$Pain = AU4 + (AU6||AU7) + (AU9||AU10) + AU43\ . \quad (12)$$

Zhang et al. [29] made extensive study to show the mapping between AUs and the targeted emotion on BP4D database, and they found that AUs $\{4, 6, 7, 9, 10\}$ are the most common in pain videos. From these results, and the available AUs annotation, we decided the beginning and the end of the pain in the videos using the following equation:

$$Pain = AU4 + (AU6 \parallel AU7) + (AU9 \parallel AU10)\ , \quad (13)$$

which states that a *physical pain* is considered as existing if AU4 and (AU6 or AU7) and (AU9 or AU10) are activated.

Based on the available AUs annotation in BP4D database, 28 subjects have been selected for the task of physical pain detection (task 6 videos). Half of these subjects (14) are used for training and half (14) for testing in the SO-SVM learning framework with the beginning and the end of pain emotion labels. Two-fold cross validation is applied here, so that every pain video out of the 28 is used both as training and testing at least once. There is no need for concatenation of GMH in these experiments, since we have long 3D videos and the pain does not start immediately according to the eliciting protocol. Two methods have been investigated in this work to model the 3D video subsequences. Results, for both the cases are reported in the following, using a window size $\omega = 6$ for deriving the linear subspaces.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

3D Landmarks-based method (baseline)

In this baseline representation, we use the 3D coordinates (x, y, z) of the 83 landmarks available in BP4D metadata as a representative feature for every 3 frames. These values are vectorized in $\mathbb{R}^n$, with $n = 83 * 3 = 249$. Then, we model every subsequence of size $\omega = 6$ as one subspace after applying k-SVD, with $k = 2$. These settings are selected empirically. Two experiments are conducted using this representation to study the effects of pose variations and of the step δ.

To evaluate the effect of pose variations on pain detection accuracy, we used the landmarks-based representation with and without pose normalization. Pose normalization is obtained by applying the inverse rotation of the 3D frame pose information given in the metadata. From Fig. 15, it results the AUC with pose normalization (0.68,0.78,0.76) are higher than without pose normalization (0.63,0.75,0.70) for $\delta = 1, 3, 6$, respectively. These results confirm that variations in landmarks position induced by pose changes are combined with those originated by pain, thus producing an overall negative effect.

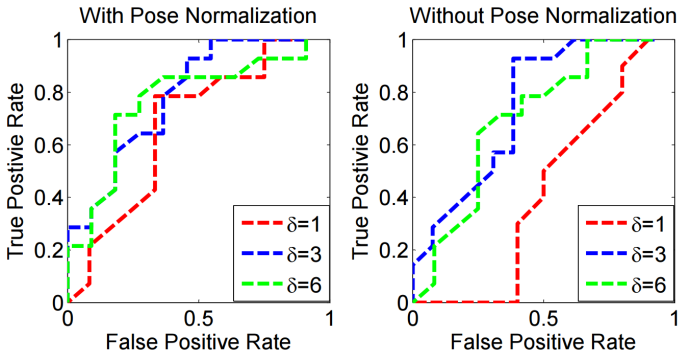

Fig. 15. ROC curve for the landmarks method. The left plots show the ROC curves after pose normalization for $\delta = \{1, 3, 6\}$, while the right plots show the performance obtained without pose normalization.

GMH extracted from curves on Grassmann manifold can be affected by noisy changes that might occur due to raw data or errors in the registration step. To investigate this aspect, we considered the effect of different smoothing levels applied to the Grassmann trajectory, which corresponds to using different values of δ. This empirical analysis is conducted using the landmarks representation with normalized pose ($\omega = 6$ and $k = 2$). Table 3 shows the AUC values for pain detection with this setting for δ from 1 to 5. The best AUC value of 0.78 is obtained for $\delta = 3$. These results show that smoothed trajectories, corresponding to $\delta > 1$, provide better performance up to a certain extent, thanks to the noise removal. However, large values of δ (e.g., $\delta = 4, 5$) affect negatively the results, since informative changes along the time can be canceled. Figure 16 illustrates the idea of trajectory smoothing.

Depth representation method

In this approach, the depth images of the face region are used instead of the landmarks. The depth image is obtained by rendering the 3D model after pose normalization, then the face region is cropped and saved as a depth image of size 100×75. The pain depth video is divided into subsequences

TABLE 3
AUC values for the landmarks method, with and without pose normalization, for $\delta = 1, 2, 3, 4, 5$

value of δ	1	2	**3**	4	5
AUC – not normalized pose	0.63	0.69	**0.75**	0.71	0.70
AUC – normalized pose	0.68	0.72	**0.78**	0.75	0.74

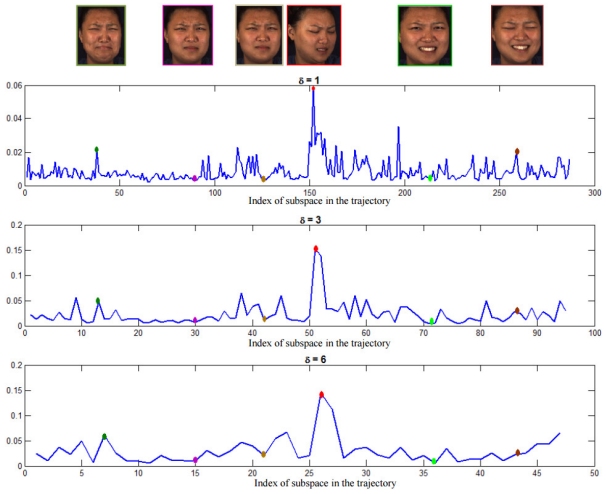

Fig. 16. The instantaneous speed (or GMH) along a trajectory computed for a depth video. Plots for $\delta = 1, 3, 6$ are reported from top to bottom.

of size $\omega = 6$, and every subsequence is modeled as one subspace by applying k-SVD, with $k = 2$ and $\delta = 3$.

Firstly, we compare the performance of the proposed pain detection framework by using two different facial representations: the landmarks, and the depth data of the face region. In both cases, the geodesic distance is used to create the GMH trajectories, with $\omega = 6$ and $k = 2$ under normalized pose. Figure 17 shows the ROC and AMOC curves for the two methods. From the ROC curve, we observe the depth representation, which captures more spatio-temporal information, also achieves better performance on pain affect detection. The AUC value obtained using depth flow reached 0.80, compared to the value of 0.78 obtained using the landmarks only. Although the overall AUC value of depth is not so much higher than the landmarks method, it is important to highlight that the TPR of the depth method is significantly higher when the FPR is less than 20%. Also, in terms of timeliness represented by AMOC curve, we can see that the depth flow method requires less seen portion of the data than the landmarks based method to give the same performance, when the FPR is less than about 20%.

The performance of the GMH is then evaluated in comparison with the proposed LDH descriptor extracted from the whole velocity vector between two subspaces along the trajectory (see Sect. 6). In both cases, we used pose normalization with $\omega = 6$, $k = 2$, and $\delta = 3$. In Fig. 18, the ROC curves on the left show the superior performance of the LDH representation over the GMH, where the AUC for LDH and GMH are 0.84 and 0.80, respectively. The AMOC curve on the right shows that the two methods are comparable, while the system receives less than 40% of the pain emotion. In particular, the LDH method achieves

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2016.2623718, IEEE Transactions on Affective Computing

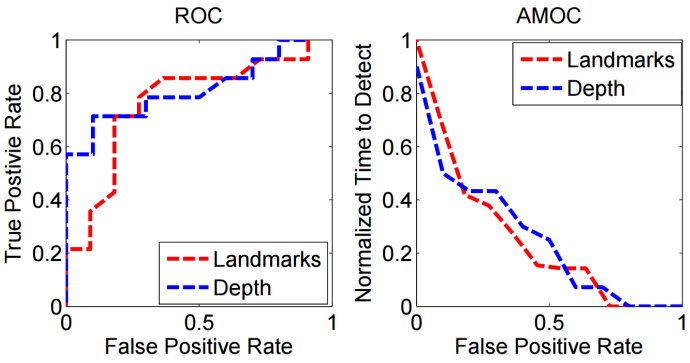

Fig. 17. ROC and AMOC curves comparing pain detection results obtained using the proposed landmarks and depth based representations.

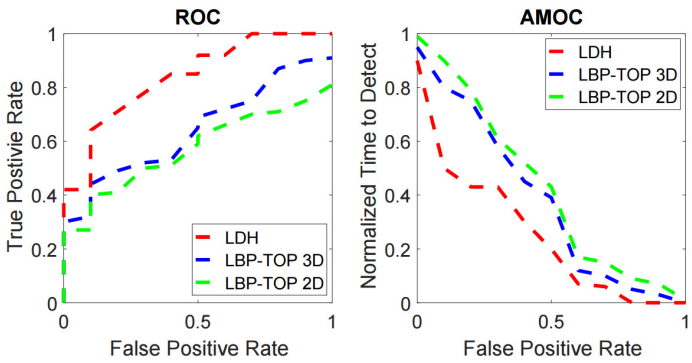

Fig. 19. ROC and AMOC curves for *pain* detection using LDH on 3D videos and LBP-TOP on 3D and 2D videos.

less FPR by seeing more frames. These results confirm the efficiency of using local coding of the temporal facial deformation through the time for pain affect detection from facial expressions. This representation outperforms the geodesic distance method, which accounts only for the speed of the deformation through the time, thus incurring in potential hiding of important local cues for detection.

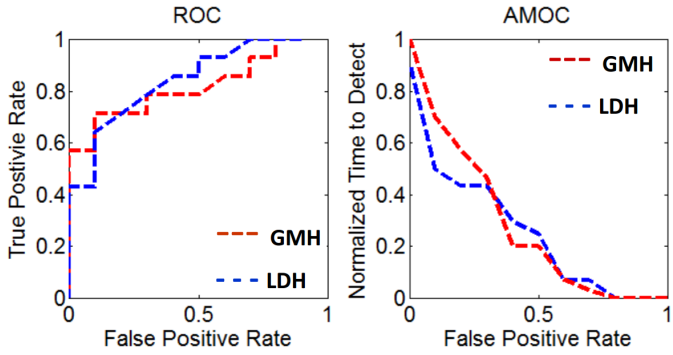

Fig. 18. ROC and AMOC curves comparing pain detection using GMH and LDH.

We evaluated the efficiency of our proposed feature also in comparison to the LBP-TOP feature applied to 3D data and 2D videos. For LBP-TOP feature extraction, we followed the same settings of Sect. 7.2. ROC and AMOC curves are presented in Fig. 19. The efficiency of the proposed method is well demonstrated from the ROC curves, where the AUC is 0.84 for LDH, 0.64 using LBP-TOP on 3D data, and 0.59 using LBP-TOP on 2D videos. The AMOC curve for LDH on 3D videos also requires less portion of the video at the same FPR than the LBP-TOP on 3D and 2D videos.

Finally, to report more robust statistical results for our depth based method, we repeated the two-fold cross validation 28 times, every time shifting by one the division between training and testing samples. The mean AUC and the standard deviation obtained are 79.8 ± 1.9. The small value of the standard deviation shows the robustness to person identity of the proposed method.

Discussion with respect to the state-of-the-art

A direct quantitative comparison of our proposed framework with the few existing pain detection or recognition works on BP4D-spontaneous database is not feasible due to different settings and problem formulation. Zhang et al. [48], addressed the problem of pain detection as a problem of AUs pain-related detection. They presented the accuracy of detecting $AU6\&7$, $AU9$ and $AU11\&12$ separately using binary edge feature at every frame and LDCRF [61] for binary classification. However, in this method detection results for individual AUs are reported, which makes the accuracy of the real pain detection unknown. Also, the problem of early pain detection is not investigated. The work of Reale et al. [62] also addresses spontaneous facial expression and AUs detection on BP4D-spontaneous dataset. In this work, the space-time *Nebula Feature* is presented, showing promising results on posed facial expression classification from 3D videos, and spontaneous AUs detection. However, pain detection is not addressed directly in this work, and the AUs detection was performed at the frame-level.

8 Conclusions and Future Work

In this paper, we have introduced a novel geometric framework for early detection of spontaneous expressions and experimented its applicability in two different scenarios: *(i)* happiness/thinking-unsure detection in depth videos of the upper part of the body acquired using Kinect-like cameras (depth-bodies); and *(ii)* physical pain detection from 3D high-resolution facial sequences (4D-faces). The key idea of our approach is to represent the stream of depth-frames as trajectories of subspaces on a Grassmann manifold. Analyzing the obtained trajectories gives rise to space-time features, where two descriptors, GMH and LDH are introduced. We have experimentally illustrated the effectiveness of the proposed framework using two datasets: the Cam3D contains spontaneous emotions and complex mental states for emotion detection from the upper part of the body, while the BP4D consists of high-resolution 4D facial sequences for physical pain affect detection. Experimental analysis of our proposed approach in comparison with 2D and 3D methods demonstrates its effectiveness. To our knowledge, this is the first work proposing early automatic detection of spontaneous emotions and pain acquired from high-resolution and low-resolution depth videos.

As future work, we will investigate advanced statistical inference techniques of partial (or full) observations using intrinsic (on the manifold) or extrinsic (e.g., fixed tangent space) methods. In addition, since our proposed

LDH descriptor provides a quantified feature vector indicating where the facial deformation happened, in which direction and its intensity, it can be used for pain-related AUs activation detection as in [48]. Furthermore, it can be used for pain intensity estimation as well. Both these aspects will be part of our future investigation. We also plan to apply the same framework on other expression detection and classification tasks on BP4D database and on cross-dataset scenarios to validate its generality and to make more detailed comparison with other detection approaches.

ACKNOWLEDGMENT

This work was supported by the Futur & Ruptures Institut Mines-Télécom Ph.D Grant to T. Alashkar and partially supported by the FUI project MAGNUM 2, the Programme d'Investissements d'Avenir (PIA), Agence Nationale pour la Recherche (grant ANR-11-EQPX-0023), European Founds for the Regional Development (Grant FEDER-Presage 41779) and the PHC Utique program for the CMCU DEFI project (N 34882WK).

REFERENCES

[1] A. Danelakis, T. Theoharis, and I. Pratikakis, "A survey on facial expression recognition in 3D video sequences," *Multimedia Tools and Applications*, vol. 74, 2014.

[2] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti, "A Grassmann framework for 4D facial shape analysis," *Pattern Recognition*, vol. 57, pp. 21–30, Sep. 2016.

[3] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Nebraska Symposium on Motivation*, vol. 19, Lincoln, NE, 1972.

[4] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *IEEE Int. Conf. on FG*, 2000.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," in *IEEE Int. Conf. on FG*, Sep 2008.

[6] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3D dynamic facial expression database," in *IEEE Int. Conf. on FG*, Sep. 2008.

[7] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. on Affective Computing,*, vol. 6, no. 1, Jan 2015.

[8] S. Eleftheriadis, O. Rudovic, and M. Pantic, "Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition," *IEEE Trans. on Image Processing*, vol. 24, Jan 2015.

[9] K. Schmidt, Z. Ambadar, J. Cohn, and L. Reed, "Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling," *Journal of Nonverbal Behavior*, vol. 3, no. 1, pp. 37–52, 2006.

[10] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, Sept 2003.

[11] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, Sep 1977.

[12] D. Watson, L. Clark, A. Tellegen, and L. Yin, "Development and validation of brief measures of positive and negative affect: the panas scales." *Journal of Personality and Social Psychology*, vol. 54, no. 6, Jun 1988.

[13] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, 2009.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, Jan 2001.

[15] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis, "Modeling naturalistic affective states via facial and vocal expressions recognition," in *Int. Conf. on Multimodal Interfaces*. ACM, 2006.

[16] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, 2013.

[17] M. Arif-Rahu and M. J. Grap, "Facial expression and pain in the critically ill non-communicative patient: State of science review," *Intensive and Critical Care Nursing*, vol. 26, pp. 343 – 352, 2010.

[18] M. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, A. Elkins, N. Tyler, P. Watson, A. Williams, M. Pantic, and N. Berthouze, "The automatic detection of chronic pain-related expression: requirements, challenges and a multimodal dataset," *IEEE Trans. on Affective Computing*, 2015.

[19] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. on Affective Computing*, vol. 4, no. 1, Jan 2013.

[20] J. Van den Stock, R. Righart, and B. de Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion*, vol. 7, Aug 2007.

[21] H. Meeren, C. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *National Academy of Sciences USA*, vol. 102, no. 45, 2005.

[22] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," *Human Computer Interaction*, vol. 28, 2012.

[23] S. Wan and J. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," *Pattern Recognition*, vol. 47, no. 5, pp. 1859–1868, 2014.

[24] M. Abd El Meguid and M. Levine, "Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers," *IEEE Trans. on Affective Computing*, vol. 5, Apr 2014.

[25] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Trans. on Affective Computing*, vol. 6, no. 1, pp. 43–55, Jan 2015.

[26] S. Mavadati, M. Mahoor, K. Bartlett, P. Trinh, and J. Cohn, "Disfa: A spontaneous facial action intensity database," *IEEE Trans. on Affective Computing*, vol. 4, no. 2, pp. 151–160, Apr 2013.

[27] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikainen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE Int. Conf. and Work. on FG*, Apr 2013, pp. 1–6.

[28] S. Aly, A. Trubanova, L. Abbott, S. White, and A. Youssef, "Vt-kfer: A kinect-based rgbd+time dataset for spontaneous and non-spontaneous facial expression recognition," in *Int. Conf. on Biometrics*, May 2015.

[29] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, 2014.

[30] M. Mahmoud, T. Baltrŭsaitis, P. Robinson, and L. Riek, "3D corpus of spontaneous complex mental states," in *Conf. on Affective Computing and Intelligent Interaction*, Oct 2011.

[31] A. Cruz, B. Bhanu, and N. Thakoor, "Vision and attention theory based sampling for continuous facial emotion recognition," *IEEE Trans. on Affective Computing*, vol. 5, no. 4, pp. 418–431, Oct 2014.

[32] Z. Zeng, Y. Fu, G. Roisman, Z. Wen, Y. Hu, and T. Huang, "One-class classification for spontaneous facial expression analysis," in *Int. Conf. on FG*, Apr 2006.

[33] I. Hupont, S. Baldassarri, and E. Cerezo, "Facial emotional classification: from a discrete perspective to a continuous emotional space," *Pattern Analysis and Applications*, vol. 16, no. 1, 2012.

[34] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[35] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[36] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Trans. on Affective Computing*, vol. 7, Jan. 2016.

[37] Y. Gaus, H. Meng, A. Jan, F. Zhang, and S. Turabzadeh, "Automatic affective dimension recognition from naturalistic facial expressions based on wavelet filtering and pls regression," in *IEEE Int. Conf. and Work. on FG*, vol. 05, May 2015, pp. 1–6.

[38] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," in *IEEE Int. Conf. and Work. on FG*, vol. 05, May 2015, pp. 1–6.

[39] K. Sikka, A. A. Ahmed, D. Diaz, M. S. Goodwin, K. D. Craig, M. S. Bartlett, and J. S. Huang, "Automated assessment of children's postoperative pain using computer vision," *Pediatrics*, 2015.

[40] K. M. Prkachin, E. Hughes, I. Schultz, P. Joy, and D. Hunt, "Real-time assessment of pain behavior during clinical assessment of low back pain patients," *Pain*, vol. 95, no. 12, 2002.

[41] A. C. Williams, "Facial expression of pain: An evolutionary account," *Behavioral and Brain Sciences*, vol. 25, 2002.

[42] C. Roy, C. Blais, D. Fiset, P. Rainville, and F. Gosselin, "Efficient information for recognizing pain in facial expressions," *European Journal of Pain*, vol. 19, no. 6, 2015.

[43] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *IEEE Int. Conf. on FG*, Mar 2011, pp. 57–64.

[44] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews, "Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database," *Image and Vision Computing*, vol. 30, 2012.

[45] S. Kaltwang, O. Rudovic, and P. M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Advances in visual computing*, ser. Lecture notes in computer science, vol. 7432. Berlin, Germany: Springer, 2012, pp. 368–377.

[46] R. Khan, A. Meyer, H. Konik, and S. Bouakaz, "Pain detection through shape and appearance features," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, Jul 2013, pp. 1–6.

[47] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," in *IEEE Int. Conf. and Workshops on FG*, Apr 2013.

[48] X. Zhang, L. Yin, and J. F. Cohn, "Three dimensional binary edge feature representation for pain expression analysis," in *IEEE Int. Conf. FG*, May 2015.

[49] Y. Sun, X. Chen, M. Rosato, and L. Yin, "Tracking vertex flow and model adaptation for 3D spatio-temporal face analysis," *IEEE Trans. on Systems, Man, and Cybernetics - Part A*, vol. 40, May 2010.

[50] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[51] M. Liu, R. Wang, Z. Huang, S. Shan, and X. Chen, "Partial least squares regression on Grassmannian manifold for emotion recognition," in *ACM Int. Conf. on Multimodal Interaction*, 2013.

[52] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Int. Conf. on Machine Learning (ICML)*, 2008.

[53] M. Hoai and F. De la Torre, "Max-margin early event detectors," *Int. Journal of Computer Vision*, vol. 107, no. 2, Feb 2014.

[54] Y. M. Lui, "Advances in matrix manifolds for computer vision," *Image Vision Computing*, vol. 30, no. 6-7, Jun 2012.

[55] G. Golub and C. Van Loan, *Matrix computations (3rd edition)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[56] N. Nguyen and Y. Guo, "Comparisons of sequence labeling algorithms and extensions," in *Int. Conf. on Machine Learning*. New York, NY, USA: ACM, 2007.

[57] Q. Rentmeesters, P.-A. Absil, P. Van Dooren, K. Gallivan, and A. Srivastava, "An efficient particle filtering technique on the Grassmann manifold," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, Mar 2010.

[58] A. Shrivastava, S. Shekhar, and V. Patel, "Unsupervised domain adaptation using parallel transport on Grassmann manifold," in *IEEE Winter Conf. on Applications of Computer Vision*, Mar 2014.

[59] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, Jun 2007.

[60] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, Oct 2008.

[61] L. P. Morency, A. Quattoni, and T. Darrell, "Efficient temporal sequence comparison and classification using gram matrix embeddings on a Riemannian manifold," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Jun 2007.

[62] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis," in *IEEE Int. Conf. and Work. on FG*, Apr 2013.

Taleb Alashkar is a Research Assistant at the Dept. of Electrical and Computer Engineering, Northeastern University, Boston, MA. He received his Ph.D degree in computer science from University of Lille 1 and Master degree in Computer Vision from University of Dijon in 2015 and 2012, respectively. He served as Principle Committee Member for ICMLA-2016 and as Reviewer for several conferences in computer vision, multimedia and machine learning. His research interests include Computer Vision, Machine Learning and Multimedia.

Boulbaba Ben Amor is an Associate Professor of computer science with the Institut Mines-Télécom/Télécom Lille and member of the CRIStAL Research Center (UMR CNRS 9189) in France. He holds an Habilitation to supervise research from University of Lille. He received his Ph.D. from Ecole Centrale de Lyon in 2006. During 2013-2014, he was a visiting research professor at Florida State University (USA). Recently, he awarded the Fulbright research grant for the year 2016-2017. He is IEEE Senior Member. He served as Area Chair for the WACV'16 conference and as Reviewer for several major conferences and Journals in computer vision. His research areas include 3D/4D shape analysis and pattern recognition.

Mohamed Daoudi is a Professor of Computer Science at the Institut Mines-Télécom/Télécom Lille and the head of Image group at CRIStAL Laboratory (UMR CNRS 9189), France. He received his Ph.D in Computer Engineering from the University of Lille1 (France) in 1993 and HDR from the University of Littoral (France) in 2000. His research interests include pattern recognition, shape analysis, computer vision and 3D object processing. He has published over 150 research papers dealing with these subjects that have appeared in the most distinguished peer-reviewed journals and conference proceedings. He is the co-author of several books including 3D Face Modeling, Analysis and Recognition, and 3D Object Processing: Compression, Indexing and Watermarking. He has been Conference Chair of SMI 2015 and several other national conferences and international workshops. He is a senior member of the IEEE, member of Association of Computing Machinery (ACM) and a fellow of the IAPR.

Stefano Berretti is an Associate Professor at the Dept. of Information Engineering and at the Media Integration and Communication Center of the University of Florence, Italy. His research interests include Pattern Recognition and 3D Computer Vision. He has published more than 120 papers in book chapters, journals and conference proceedings. He is Information Director of the ACM Transactions on Multimedia Computing, Communications, and Applications.